LA-UR- LA-UR- 02-1434

Approved for public release; distribution is unlimited.

Title: Orthogonal Arrays for Computer Experiments to Assess ~~Import~~ Inputs
important

Author(s): L.M. Moore, D-1
M.D. McKay, D-1

Submitted to: 6th International Conference on Probabilistic Safety Assessment and Management, Puerto Rico

Los Alamos
NATIONAL LABORATORY

Orthogonal Arrays for Computer Experiments to Assess Important Inputs

L. M. Moore and M. D. McKay

Los Alamos National Laboratory, Los Alamos, NM 87545

ABSTRACT

The topic of this paper is experiment planning, particularly fractional factorial designs or orthogonal arrays, for computer experiments to assess important inputs. The work presented in the paper is motivated by considering a non-stochastic computer simulation which has many inputs and which can, in a reasonable period of time, be run thousands of times. With many inputs, information that allows focus on a subset of important inputs is valuable. The characterization of "importance" is expected to follow suggestions in McKay (1995) or McKay, et al. (1992). This analysis approach leads to considering factorial experiment designs. Inputs are associated with a finite number of discrete values, referred to as levels, so if each input has K levels and there are p inputs then there are K^P possible distinct runs which constitute the K^P factorial design space. The suggested size of p has been 35 to 50 so that even with K=2 the complete 2^P factorial design space would not be run. Further, it is expected that the complexity of the simulation code and discrete levels possibly associated with equi-probable intervals from the input distribution make it desirable to consider more than 2 level inputs. Inputs levels of 5 and 7 have been investigated. In this paper, orthogonal array experiment designs, which are subsets of factorial designs also referred to as fractional factorial designs, are suggested as candidate experiments which provide meaningful basis for calculating and comparing R^2 across subsets of inputs.

KEYWORDS

Computer experiments, experiment design, fractional factorial design, orthogonal arrays, correlation coefficient

1. INTRODUCTION

The topic of this paper is experiment planning, particularly fractional factorial designs or orthogonal arrays, for computer experiments to assess important inputs. The work presented in the paper is motivated by considering a non-stochastic computer simulation which has many inputs and which can, in a reasonable period of time, be run thousands of times. With many inputs, information that allows focus on a subset of important inputs is valuable. The characterization of "importance" is expected to follow suggestions in McKay (1995) or McKay, et al. (1992). One approach to screening inputs to a computer code is by identifying inputs as important according to the level of output uncertainty induced by the distribution on the inputs. If Y=h(X) denotes the calculated output, a scalar for simplicity, at input vector, X, of length p, then X_s, a subset of s<p of the p input variables, is assessed important if it has a large value of R^2, an estimate of the correlation coefficient associated with the goodness of fit to Y of an analysis of variance model based on X_s. Comparison of R^2 across subsets of the p inputs is the basis for identifying subsets of important inputs.

Based on these analysis considerations, experiment plans are required that will yield useful data for obtaining meaningful R^2 values for subsets of the possible inputs. Consider the following formula for R^2 based on a subset of inputs X_s:

$$R^2(X_s) = \frac{\sum_{i \in X_s} \sum_j (y_{i\cdot} - y_{\cdot\cdot})^2}{\sum_{i \in X_s} \sum_j (y_{ij} - y_{\cdot\cdot})^2}$$

where the subscript i varies over distinct values of the s inputs identified in X_s, the subscript j varies over "replicate" experiments corresponding to a fixed value of the inputs X_s, and the "dot" subscript indicates the standard average. "Replicate" is in guotes since no true replicates are done. The computer simulation output is non-stochastic in that the output is fully determined by specification of the input with no variation in output for repeated runs of the code for identical input. Variation in the output is induced solely by variation in the inputs. The (p-s) inputs identified by X-X_s may differ while X_s is fixed to obtain pseudo-replicate runs. It is clear that the value for $y_{i\cdot}$ will be identically y_{ij} if there are no pseudo-replicate runs. If this is the case for every value of the inputs identified by X_s, then R^2 will have a value identically 1. Otherwise, R^2 is between 0 and 1. This reasoning leads to considering experiment designs such that, for subsets of inputs of a specified size s<p, a sampling of values for that subset of inputs is required such that "replicates" determined by a sample of values for the remaining inputs occur, for at least one of the values of the subset of inputs. This is a property of factorial experiment designs.

Factorial experiments are experiments for inputs, called factors, with a finite number of discrete values, referred to as levels, so if each input has K levels and there are p inputs then there are K^P possible distinct runs referred to as the K^P factorial design space. The K levels could be associated with K equal probability content intervals for a continuous input. If the experiment plan consisted of the entire K^P factorial design space, then for each pair of inputs (subsets of size 2) there are K^2 values (levels) with K^{P-2} "replicates" for each value. Obviously this extends to subsets of inputs of size s in the obvious way. For relatively moderate K and even small sizes for p the full product space of possible experiment runs quickly becomes unmanageably large, even given the ability to run the simulation code thousands of times. The suggested size of p has been 35 to 50 so that even with K=2 the complete 2^P factorial design space would not be run. Further, it is expected that the complexity of the simulation code and some already specified discrete levels of inputs make it desirable to consider more than 2 level inputs, and inputs levels of 5 and 7 have been suggested. In this paper, orthogonal array experiment designs, which are subsets of factorial designs also referred to as fractional factorial designs, are suggested as candidate experiments which provide meaningful basis for calculating and comparing R^2 for input subsets of size 1 or 2 and have significantly fewer runs than full factorial design spaces.

In Section 2, orthogonal arrays are defined and described in more detail. In Section 3, some specific candidate orthogonal arrays are presented for 7 or 5 level factors. Concluding remarks are made in Section 4.

2. ORTHOGONAL ARRAYS

Hedayat, et al (1999) provides a good reference on orthogonal arrays. For K levels identified by elements in the set $L=\{0,1,2,\ldots,k-1\}$, an $N\times p$ array X with entries from L is an orthogonal array with K levels, strength t ($0\leq t\leq p$) and index λ if every $N\times t$ subarray of X contains each t-tuple based on L exactly λ times as a row. An array with parameters N, p, k, and t is denoted OA(N,p,k,t). From this definition, a strength t orthogonal array with index λ is a set of p-dimensional factorial design points such that if one considers any t-dimensional projection then every point in the K^t factorial design space is replicated λ times. Likewise, any projection of dimension smaller than t, say s<t, consist of $\lambda*K^{(t-s)}$ replicates of the K^S factorial design space.

Orthogonal arrays occur in statistical applications as fractional factorial designs. A full K^P factorial design space is itself an $OA(K^P,p,K,p)$ with index unity, that is $\lambda=1$. There are many textbooks on statistical experiment design including Raktoe, et al (1981) and John (1971) which are particularly useful for fractional factorial experiment design. In a strict sense, fractional factorial designs may be any subset of the full factorial design space but often this terminology is reserved for subsets that form an orthogonal array. For K prime, fractions of resolution III, IV and V defined in John (1971) or Raktoe, et al (1981) correspond to orthogonal arrays of strength 2, 3, and 4 respectively for which "replicate" runs occur for X_s including all values in the K^S grid, where s<t and t is the strength of the array.

In Raktoe, et al (1981), a fraction of the factorial design space, or orthogonal array, is defined by either a subspace or a coset of a subspace of the design space considered as a vector space of p dimensions over the Galois field generated by the symbols $L=\{0,1,2,\ldots,k-1\}$. Specifically, the K^P factorial design space is associated with the set of p-tuples $(x_1,x_2,x_3,\ldots x_p)\in\{0,1,2,\ldots,k-1\}^P$ and under component-wise arithmetic modulus K this set forms a vector space. A fraction of size K^{p-a} for $1\leq a\leq(p-1)$ is defined by the design space elements that satisfy a set of a independent and consistent equations in the vector space elements. To illustrate and dispense with the clear extensions to the general case, consider K=2 and p=5. The full 2^5 factorial design space is

$$
\begin{aligned}
\{0,1\}^5 = \{ \\
&(0,0,0,0,0),(1,0,0,0,0),(0,1,0,0,0),(1,1,0,0,0),\\
&(0,0,1,0,0),(1,0,1,0,0),(0,1,1,0,0),(1,1,1,0,0),\\
&(0,0,0,1,0),(1,0,0,1,0),(0,1,0,1,0),(1,1,0,1,0),\\
&(0,0,1,1,0),(1,0,1,1,0),(0,1,1,1,0),(1,1,1,1,0),\\
&(0,0,0,0,1),(1,0,0,0,1),(0,1,0,0,1),(1,1,0,0,1),\\
&(0,0,1,0,1),(1,0,1,0,1),(0,1,1,0,1),(1,1,1,0,1),\\
&(0,0,0,1,1),(1,0,0,1,1),(0,1,0,1,1),(1,1,0,1,1),\\
&(0,0,1,1,1),(1,0,1,1,1),(0,1,1,1,1),(1,1,1,1,1)\}
\end{aligned}
$$

This listing is arranged in columns according to the 4 values of the first two inputs, (x_1,x_2), so that it is readily seen that there are 8 "replicates" on each of these 4 pair values. There are likewise 8 "replicates" on each of the 4 pair values for the other 9 input subsets of size 2.

One possible fraction of size $2^{(5-2)}$ is defined by the following equations in the first two inputs:

$$x_1 = 0 \bmod 2$$
$$x_2 = 0 \bmod 2.$$

This fraction is in fact the subspace consisting of the first column in the complete factorial design listing above. It is clear that not every pair value occurs for each variable pair, although there are "replicates" on each pair value that occurs. Also, if we limit consideration to the three inputs (x_3,x_4,x_5) then the array is strength 3.

Another fraction that is additionally a strength 2 orthogonal array is defined by the following equations:

$$x_1 + x_2 + x_3 = 0 \bmod 2$$
$$x_3 + x_4 + x_5 = 0 \bmod 2.$$

These two equations additionally imply:

$$x_1 + x_2 + x_4 + x_5 = 0 \bmod 2.$$

The fraction identified consists of the following design points:

$$\{(0,0,0,0,0),(1,1,0,0,0),(1,0,1,1,0),(0,1,1,1,0),$$
$$(1,0,1,0,1),(0,1,1,0,1),(0,0,0,1,1),(1,1,0,1,1)\}.$$

It is easily verified that each pair of inputs has 2 "replicates" each on 4 values.

The strength of the array defined in this way is determined by the "shortest" among the set of equations satisfied by the components of the design points. In general, an equation satisfied by the components of the design points $(x_1,x_2,x_3,\ldots,x_p)$ may be written as follows:

$$a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + \ldots + a_p * x_p = 0 \bmod k$$

where $(a_1,a_2,a_3,\ldots a_p)$ has components in $L=\{0,1,2,\ldots,k-1\}$. The length of an equation is the number of non-zero elements of $(a_1,a_2,a_3,\ldots a_p)$. The strength of an array defined by such a group of equations, generated by a independent and consistent equations, is one less than the minimum length of equations in the group. In the first illustration above, the minimum length equations are length 1 so the resulting array has no strength, and in fact would not be considered an orthogonal array. The second example has minimum generator equation length 3 and the resulting fraction is an orthogonal array of strength 2.

3. SPECIFIC ORTHOGONAL ARRAYS FOR K=5 OR K=7

Hedayat, et al (1999) provide various bounds relating the parameters of an OA(N,p,k,t). Although not specifically mentioned above, the modulo k arithmetic generally has undesirable performance for k not prime, so k is assumed prime. There are other intricate combinatorial properties that exist for orthogonal arrays that warrant further investigation but in the following a few simple ideas are exploited to generate and describe some particular cases of orthogonal arrays for K=7 and 5.

K=7:

Rao (1946) constructed a family of arrays OA(K^n,(K^n-1)/(K-1),K,2), K prime, referred to in Hedayat, et al (1999) as Rao-Hamming type since they have the same construction as Hamming (1950) error-correcting codes. Denote the columns of the full K^n factorial design by z_i, i=1,…,n. Then the Rao-Hamming construction defines an OA(K^n,(K^n-1)/(K-1),K,2) as all columns of the form:

$$a_1 * z_1 + a_2 * z_2 + a_3 * x_3 + \ldots + a_n * z_n \bmod k$$

with (a_1,a_2,a_3,…a_n) having components in L={0,1,2,…,k-1} such that not all are 0 and the first nonzero a_i has value 1. There are precisely (K^n-1)/(K-1) such columns. In fact, this is the maximum number of columns possible for a strength 2 orthogonal array with K^n rows.

Applying the Rao-Hamming construction for 7 level factors provides a 343×57 strength 2 orthogonal array. Take the first 3 columns, (x_1,x_2,x_3), to be the 7^3 full factorial. Then the remaining 54 columns may be taken as follows:

$$x_4 = 1*x_1 + 1*x_2 + 0*x_3 \bmod 7, x_5 = 2*x_1 + 1*x_2 + 0*x_3 \bmod 7, \ldots, x_9 = 6*x_1 + 1*x_2 + 0*x_3 \bmod 7,$$
$$x_{10} = 1*x_1 + 0*x_2 + 1*x_3 \bmod 7, x_{11} = 2*x_1 + 0*x_2 + 1*x_3 \bmod 7, \ldots, x_{15} = 6*x_1 + 0*x_2 + 1*x_3 \bmod 7,$$
$$x_{16} = 0*x_1 + 1*x_2 + 1*x_3 \bmod 7, x_{17} = 1*x_1 + 1*x_2 + 1*x_3 \bmod 7, \ldots, x_{22} = 6*x_1 + 1*x_2 + 1*x_3 \bmod 7,$$
$$x_{23} = 0*x_1 + 2*x_2 + 1*x_3 \bmod 7, x_{24} = 1*x_1 + 2*x_2 + 1*x_3 \bmod 7, \ldots, x_{29} = 6*x_1 + 2*x_2 + 1*x_3 \bmod 7,$$
$$x_{30} = 0*x_1 + 3*x_2 + 1*x_3 \bmod 7, x_{31} = 1*x_1 + 3*x_2 + 1*x_3 \bmod 7, \ldots, x_{36} = 6*x_1 + 3*x_2 + 1*x_3 \bmod 7,$$
$$x_{37} = 0*x_1 + 4*x_2 + 1*x_3 \bmod 7, x_{38} = 1*x_1 + 4*x_2 + 1*x_3 \bmod 7, \ldots, x_{43} = 6*x_1 + 4*x_2 + 1*x_3 \bmod 7,$$
$$x_{44} = 0*x_1 + 5*x_2 + 1*x_3 \bmod 7, x_{45} = 1*x_1 + 5*x_2 + 1*x_3 \bmod 7, \ldots, x_{50} = 6*x_1 + 5*x_2 + 1*x_3 \bmod 7,$$
$$x_{51} = 0*x_1 + 6*x_2 + 1*x_3 \bmod 7, x_{52} = 1*x_1 + 6*x_2 + 1*x_3 \bmod 7, \ldots, x_{57} = 6*x_1 + 6*x_2 + 1*x_3 \bmod 7.$$

It is straightforward to write a short program to generate the 343 runs defined by these equations and computationally check that it is in fact a strength 2 orthogonal array. Since there are equations of length 3 among the 54 generator equations, the array would be at most strength 2 (less if an equation implied by combinations of the generator equations happened to have shorter length). The strength 2 property means that for any pair of inputs (columns), all of the 49 possible pair values are "replicated" 7 times. There are 57 choose 2, or 1596 pairs of inputs. Additional properties of this design include that for the 57 choose 3 columns (29260 triple inputs), except for 3192 cases, each of the 343 possible triple values occurs precisely once. For those triples of inputs where not all values occur, if one considers any two columns and one of the 49 fixed pair values then the value of the third column in the triple is fixed to one value in L={1,2,3,4,5,6} for those 7 cases of fixed pair value. This property is referred to as aliasing of pairs of columns with single columns. There are 6 single columns aliased with each pair of columns.

The impact on analysis of aliasing (or confounding) needs to be characterized. In terms of a standard statistical model associated with a factorial design based on 57 inputs, main effects are estimable but biased by any possibly active two factor (and higher) interactions with which they may be aliased. Although the $R^2(x_i,x_j)$ values can be calculated and are not degenerate to 1, assessment of a most important pair of inputs may similarly be biased by a most important single input that is not actually one of the best pair.

The maximum number of columns possible for a 343 row (run) orthogonal array of strength 2 is 57. The Introduction stated 35 to 50 7-level factors are expected. It is clear that any subset $p \le 57$ of the columns is also an OA(343,p,7,2). Generally, an experimenter may pick any of the 57 columns if fewer than 57 factors are required. However, all subsets may not be equal and criteria for selecting fewer columns exist and could be applied. One consideration is how many columns of an

$OA(7^3,57,7,2)$ might actually be strength 3? Theorem 3.1 of Hedayat, et al (1999) indicates that an $OA(K^3,K+1,K,3)$ exists whenever $K \geq (t-1) \geq 0$. So for K=7, there are 8 columns of the OA(343,57,7,2) that actually form an OA(343,8,7,3).

K=5:

Hedayat, et al (1999) has several tables of parameters for orthogonal array that exist and references for construction methods. With few exceptions, most of the arrays in these tables are strength 2 or consider only 2 or 3 level factors. Computer experiments, especially those with many inputs and the possibility of thousands of runs, allow the consideration of strength 3 or 4 arrays. Table 12.6(a) in Hedayat, et al (1999) indicate the existence of an $OA(K^4,K^2+1,K,3)$, or for K=5 an OA(625,26,5,3). For 3 level factors, Table 12.6(e) lists the existence of OA(81,10,3,3) and OA(243,11,3,4). The construction references there are based on finite geometries or ternary error-correcting code. Here, an approach to constructions and descriptions based on the modulus K equations is implemented.

The Rao-Hamming construction for 5 level factors provides a set of 152 modulus 5 equations that define a 625×156 strength 2 orthogonal array similarly to the case above for K=7. In this case, take the first 4 columns, (x_1,x_2,x_3,x_4), to be the 5^4 full factorial. Then the remaining columns may be taken as follows:

length 3 generators (there are (4choose2)*4=6*4=24)

$x_5 = 0*x_1 + 0*x_2 + 1*x_3 + 1*x_4 \bmod 5, ..., x_8 = 0*x_1 + 0*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_9 = 0*x_1 + 1*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{12} = 0*x_1 + 1*x_2 + 0*x_3 + 4*x_4 \bmod 5,$

$x_{13} = 0*x_1 + 1*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{16} = 0*x_1 + 1*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{17} = 1*x_1 + 0*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{20} = 1*x_1 + 0*x_2 + 0*x_3 + 4*x_4 \bmod 5,$

$x_{21} = 1*x_1 + 0*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{24} = 1*x_1 + 0*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{25} = 1*x_1 + 1*x_2 + 0*x_3 + 0*x_4 \bmod 5, ..., x_{28} = 1*x_1 + 4*x_2 + 0*x_3 + 0*x_4 \bmod 5,$

length 4 generators (there are 4*4*4=64)

$x_{29} = 0*x_1 + 1*x_2 + 1*x_3 + 1*x_4 \bmod 5, ..., x_{32} = 0*x_1 + 1*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{33} = 0*x_1 + 1*x_2 + 2*x_3 + 1*x_4 \bmod 5, ..., x_{36} = 0*x_1 + 1*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{37} = 0*x_1 + 1*x_2 + 3*x_3 + 1*x_4 \bmod 5, ..., x_{40} = 0*x_1 + 1*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{41} = 0*x_1 + 1*x_2 + 4*x_3 + 1*x_4 \bmod 5, ..., x_{44} = 0*x_1 + 1*x_2 + 4*x_3 + 4*x_4 \bmod 5,$

$x_{45} = 1*x_1 + 0*x_2 + 1*x_3 + 1*x_4 \bmod 5, ..., x_{48} = 1*x_1 + 0*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{49} = 1*x_1 + 0*x_2 + 2*x_3 + 1*x_4 \bmod 5,, ..., x_{52} = 1*x_1 + 0*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{53} = 1*x_1 + 0*x_2 + 3*x_3 + 1*x_4 \bmod 5, ..., x_{56} = 1*x_1 + 0*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{57} = 1*x_1 + 0*x_2 + 4*x_3 + 1*x_4 \bmod 5, ..., x_{60} = 1*x_1 + 0*x_2 + 4*x_3 + 4*x_4 \bmod 5,$

$x_{61} = 1*x_1 + 1*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{64} = 1*x_1 + 1*x_2 + 0*x_3 + 4*x_4 \bmod 5,$

$x_{65} = 1*x_1 + 2*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{68} = 1*x_1 + 2*x_2 + 0*x_3 + 4*x_4 \bmod 5,$

$x_{69} = 1*x_1 + 3*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{72} = 1*x_1 + 3*x_2 + 0*x_3 + 4*x_4 \bmod 5,$

$x_{73} = 1*x_1 + 4*x_2 + 0*x_3 + 1*x_4 \bmod 5, ..., x_{76} = 1*x_1 + 4*x_2 + 0*x_3 + 4*x_4 \bmod 5$

$x_{77} = 1*x_1 + 1*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{80} = 1*x_1 + 1*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{81} = 1*x_1 + 2*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{84} = 1*x_1 + 2*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{85} = 1*x_1 + 3*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{88} = 1*x_1 + 3*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{89} = 1*x_1 + 4*x_2 + 1*x_3 + 0*x_4 \bmod 5, ..., x_{92} = 1*x_1 + 4*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

and

length 5 generators (there are 4*4*4=64)

$x_{93} = 1*x_1 + 1*x_2 + 1*x_3 + 1*x_4 \bmod 5, \ldots, x_{96} = 1*x_1 + 1*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{97} = 1*x_1 + 1*x_2 + 2*x_3 + 1*x_4 \bmod 5, \ldots, x_{100} = 1*x_1 + 1*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{101} = 1*x_1 + 1*x_2 + 3*x_3 + 1*x_4 \bmod 5, \ldots, x_{104} = 1*x_1 + 1*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{105} = 1*x_1 + 1*x_2 + 4*x_3 + 1*x_4 \bmod 5, \ldots, x_{108} = 0*x_1 + 1*x_2 + 4*x_3 + 4*x_4 \bmod 5,$

$x_{109} = 1*x_1 + 2*x_2 + 1*x_3 + 1*x_4 \bmod 5, \ldots, x_{112} = 1*x_1 + 2*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{113} = 1*x_1 + 2*x_2 + 2*x_3 + 1*x_4 \bmod 5, \ldots, x_{116} = 1*x_1 + 2*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{117} = 1*x_1 + 2*x_2 + 3*x_3 + 1*x_4 \bmod 5, \ldots, x_{120} = 1*x_1 + 2*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{121} = 1*x_1 + 2*x_2 + 4*x_3 + 1*x_4 \bmod 5, \ldots, x_{124} = 1*x_1 + 2*x_2 + 4*x_3 + 4*x_4 \bmod 5,$

$x_{125} = 1*x_1 + 3*x_2 + 1*x_3 + 1*x_4 \bmod 5, \ldots, x_{128} = 1*x_1 + 3*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{129} = 1*x_1 + 3*x_2 + 2*x_3 + 1*x_4 \bmod 5, \ldots, x_{132} = 1*x_1 + 3*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{133} = 1*x_1 + 3*x_2 + 3*x_3 + 1*x_4 \bmod 5, \ldots, x_{136} = 1*x_1 + 3*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{137} = 1*x_1 + 3*x_2 + 4*x_3 + 1*x_4 \bmod 5, \ldots, x_{140} = 1*x_1 + 3*x_2 + 4*x_3 + 4*x_4 \bmod 5$

$x_{141} = 1*x_1 + 4*x_2 + 1*x_3 + 1*x_4 \bmod 5, \ldots, x_{144} = 1*x_1 + 4*x_2 + 1*x_3 + 4*x_4 \bmod 5,$

$x_{145} = 1*x_1 + 4*x_2 + 2*x_3 + 1*x_4 \bmod 5, \ldots, x_{148} = 1*x_1 + 4*x_2 + 2*x_3 + 4*x_4 \bmod 5,$

$x_{149} = 1*x_1 + 4*x_2 + 3*x_3 + 1*x_4 \bmod 5, \ldots, x_{152} = 1*x_1 + 4*x_2 + 3*x_3 + 4*x_4 \bmod 5,$

$x_{153} = 1*x_1 + 4*x_2 + 4*x_3 + 1*x_4 \bmod 5, \ldots, x_{156} = 1*x_1 + 4*x_2 + 4*x_3 + 4*x_4 \bmod 5.$

Again it is straightforward to write a short program to generate the 625×156 strength 2 array for 5 level factors. Next consider whether there is a subset of the 156 columns that form a strength 3 array. Fixing the first four columns, an additional 22 columns could be selected from the 152 columns and tested for the strength 3 property. Recalling that for an array to be strength 3 the minimum length of any defining equation must be 4, the search can be restricted to selecting 22 from the 128 columns defined above with generator equation length 4 or 5. Of course, an exhaustive search of all 128 choose 22 columns would consume an impossible amount of computer time, but a random search fairly quickly yielded a case of an OA(625,26,5,3). The subset of 26 columns include the first 4 and: x_{35}, x_{37}, x_{42}, x_{47}, x_{52}, x_{54}, x_{65}, x_{70}, x_{76}, x_{84}, x_{87}, x_{89}, x_{94}, x_{97}, x_{104}, x_{107}, x_{114}, x_{119}, x_{125}, x_{140}, x_{147}, and x_{153}. The generating equations are:

length 4 generators (12)

$x_{35} = 0*x_1 + 1*x_2 + 2*x_3 + 3*x_4 \bmod 5, x_{37} = 0*x_1 + 1*x_2 + 3*x_3 + 1*x_4 \bmod 5,$

$x_{42} = 0*x_1 + 1*x_2 + 4*x_3 + 2*x_4 \bmod 5, x_{47} = 1*x_1 + 0*x_2 + 1*x_3 + 3*x_4 \bmod 5,$

$x_{52} = 1*x_1 + 0*x_2 + 2*x_3 + 4*x_4 \bmod 5, x_{54} = 1*x_1 + 0*x_2 + 3*x_3 + 2*x_4 \bmod 5,$

$x_{65} = 1*x_1 + 2*x_2 + 0*x_3 + 1*x_4 \bmod 5, x_{70} = 1*x_1 + 3*x_2 + 0*x_3 + 2*x_4 \bmod 5,$

$x_{76} = 1*x_1 + 4*x_2 + 0*x_3 + 4*x_4 \bmod 5, x_{84} = 1*x_1 + 2*x_2 + 4*x_3 + 0*x_4 \bmod 5,$

$x_{87} = 1*x_1 + 3*x_2 + 3*x_3 + 0*x_4 \bmod 5, x_{89} = 1*x_1 + 4*x_2 + 1*x_3 + 0*x_4 \bmod 5,$

length 5 generators (10)

$x_{94} = 1*x_1 + 1*x_2 + 1*x_3 + 2*x_4 \bmod 5, x_{97} = 1*x_1 + 1*x_2 + 2*x_3 + 1*x_4 \bmod 5,$

$x_{104} = 1*x_1 + 1*x_2 + 3*x_3 + 4*x_4 \bmod 5, x_{107} = 1*x_1 + 1*x_2 + 4*x_3 + 3*x_4 \bmod 5,$

$x_{114} = 1*x_1 + 2*x_2 + 2*x_3 + 2*x_4 \bmod 5, x_{119} = 1*x_1 + 2*x_2 + 3*x_3 + 3*x_4 \bmod 5,$

$x_{125} = 1*x_1 + 3*x_2 + 1*x_3 + 1*x_4 \bmod 5, x_{140} = 1*x_1 + 3*x_2 + 4*x_3 + 4*x_4 \bmod 5,$

$x_{147} = 1*x_1 + 4*x_2 + 2*x_3 + 3*x_4 \bmod 5, x_{153} = 1*x_1 + 4*x_2 + 4*x_3 + 1*x_4 \bmod 5.$

The random search approach finds this case but it leaves open the possibility that there are other cases, perhaps with more length 5 generators (or fewer length 4 generators). (This property is related to minimum aberration design.)

Again, the strength 3 property means that for any triple of inputs (columns), all of the 125 possible triple values are "replicated" 5 times. There are 26 choose 3, or 2600 sets of 3 inputs. The length 4 generator equations make it clear that for some input subsets of size 4 not every (625) combination occurs. Again, this property indicates aliasing of pairs of columns with other pairs of columns. The result is that some pairs of inputs have identical values for R^2. In this case, if the R^2 value is large then judgement must be made as to which of the aliased pairs of inputs is likely to be important and there is no way of distinguishing based on the simulation experiment.

The maximum number of columns possible for a 625 row (run) orthogonal array in 5 level inputs of strength 2 is 156. We have demonstrated a 26 column strength 3 array with 625 rows. Inequalities due to Rao (1947) provide an upper bound of 32 on p for an $OA(5^4=625,p,5,3)$, but results due to Bose (1947) give p=26 maximum for arrays generated by the modulus k equations considered here. Theorem 3.1 of Hedayat, et al (1999) indicates that an $OA(K^t,K+1,K,t)$ exists whenever $K \geq (t-1) \geq 0$. For K=5, there are 6 columns of the above OA(625,26,5,3) that actually form an OA(625,6,5,4). These columns are (x_1,x_2,x_3, x_4)) and

$$x_{94} = 1*x_1 + 1*x_2 + 1*x_3 + 2*x_4 \bmod 5 \text{ and}$$
$$x_{119} = 1*x_1 + 2*x_2 + 3*x_3 + 3*x_4 \bmod 5.$$

4. CONCLUSIONS

Orthogonal arrays, or fractional factorial designs, are suggested as candidates for computer experiments where assessment of important inputs might be based on the sample correlation coefficient, R^2. For such designs the pseudo-replication inherent in the balance and grid coverage for lower dimensional input space allow meaningful calculation of R^2. Alternative design strategies, such as replicated Latin Hypercube sampling (McKay, Conover, and Beckman (1979)) have been used as well as the specific designs mentioned here and the orthogonal arrays seem to perform well. Further investigation is continuing.

REFERENCES

Bose, R. C. (1947). Mathematical theory of the symmetrical factorial design. *Sankhya*, 8, pp. 107-166.

Hamming, R. W. (1950). Error-detecting and error correcting codes. *Bell Syst. Tech. J.*, 29, pp. 147-160.

Hedayat, A. S., N. J. A. Sloane, and John Stufken (1999). **Orthogonal Arrays**. Springer-Verlag New York, Inc.

John, P. W. M. (1971). **Statistical Design and Analysis of Experiments**. The Macmillan Company, New York.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21, No. 2, pp. 239-245.

McKay, M. D., R. J. Beckman, L. M. Moore, and R. R. Picard (1992). An alternative view of sensitivity in the analysis of computer codes. *Proceedings of the American Statistical Association Section on Physical and Engineering Sciences*, (Boston, Massachusetts, August 9-13).

McKay, M. D. (1995). Evaluating Prediction Uncertainty. NUREG/CR-6311, LA-12915-MS.

Raktoe, B. L., A. Hedayat, and W. T. Federer (1981). **Factorial Designs**. John Wiley & Sons, Inc.

Rao, C. R. (1946). Hypercubes of strength *d* leading to confounded designs in factorial experiments. *Bull. Calcutta Math. Soc.*, 38, pp. 67-78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *J. Royal Statist. Soc. (Suppl.)*, 9, pp. 128-139.